



Technische Universität Berlin
Institut für Mathematik

**A Matlab Toolbox for the Regularization of Descriptor Systems
Arising from Generalized Realization Procedures**

A. Binder V. Mehrmann A. Miedlar P. Schulze

Preprint 24-2015

**Preprint-Reihe des Instituts für Mathematik
Technische Universität Berlin
<http://www.math.tu-berlin.de/preprints>**

A **Matlab** Toolbox for the Regularization of Descriptor Systems Arising from Generalized Realization Procedures

A. Binder* V. Mehrmann* A. Miedlar† P. Schulze*

January 7, 2016

In this report we introduce a **Matlab** toolbox for the regularization of descriptor systems. We apply it, in particular, for systems resulting from the generalized realization procedure of [16], which generates, via rational interpolation techniques, a linear *descriptor system* from interpolation data. The resulting system needs to be regularized to make it feasible for the use in simulation, optimization, and control. This process is called *regularization*.

1 Descriptor Systems

We follow the notation and the basic concepts of [5]. A linear *descriptor system* is of the form

$$E\dot{x} = Ax + Bu, \tag{1a}$$

$$y = Cx + Du, \tag{1b}$$

where $E, A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,m}$, $C \in \mathbb{R}^{p,n}$, $D \in \mathbb{R}^{p,m}$, and $\dot{x} = dx/dt$. The response of a descriptor system can be described in terms of the eigenvalues of the matrix pencil $\alpha E - \beta A$, which is said to be *regular* if $\det(\alpha E - \beta A) \neq 0$ for some $(\alpha, \beta) \in \mathbb{C}^2$. For regular pencils, *generalized eigenvalues* are the pairs $(\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}$, for which $\det(\alpha E - \beta A) = 0$. If $\beta \neq 0$, then the pair represents the finite eigenvalue $\lambda = \alpha/\beta$. If $\beta = 0$, then (α, β) represents an infinite eigenvalue.

In frequency domain, for zero initial conditions $x(t_0) = 0$ and a regular pencil $\alpha E - \beta A$, there exists the rational *transfer function*

$$H(s) = C(sE - A)^{-1}B + D, \tag{2}$$

which maps Laplace transforms of the input functions u to the Laplace transforms of the corresponding output functions y . A finite eigenvalue $\lambda = \alpha/\beta$ is a pole of the transfer function of the descriptor system (1).

In the following we denote a matrix with orthonormal columns spanning the right nullspace of the matrix M by $S_\infty(M)$ and a matrix with orthonormal columns spanning the left nullspace of M by $T_\infty(M)$. These matrices are not uniquely determined although the spaces are, but for ease of notation, we speak of these matrices as the corresponding spaces.

For regular pencils the solution of the system equations can be characterized in terms of the Weierstraß Canonical Form (WCF), [10].

*Institut für Mathematik, TU Berlin, Germany, {binder,pschulze,mehrmann}@math.tu-berlin.de

†University of Minnesota, Minneapolis, USA amiedlar@umn.edu.

Theorem 1.1. Weierstraß Canonical Form *If $\alpha E - \beta A$ is a regular pencil, then there exist nonsingular matrices $X = [X_r, X_\infty] \in \mathbb{R}^{n,n}$ and $Y = [Y_r, Y_\infty] \in \mathbb{R}^{n,n}$ for which*

$$Y^T E X = \begin{bmatrix} Y_r^T \\ Y_\infty^T \end{bmatrix} E \begin{bmatrix} X_r & X_\infty \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad (3)$$

and

$$Y^T A X = \begin{bmatrix} Y_r^T \\ Y_\infty^T \end{bmatrix} A \begin{bmatrix} X_r & X_\infty \end{bmatrix} = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}, \quad (4)$$

where J is a matrix in Jordan canonical form whose diagonal elements are the finite eigenvalues of the pencil and N is a nilpotent matrix, also in Jordan form. J and N are unique up to permutation of Jordan blocks.

The index ν of the pencil $\alpha E - \beta A$ is the index of nilpotency of the nilpotent matrix N in (3). By convention, if E is nonsingular, the pencil is said to be of index zero. A descriptor system is regular and of index at most one if and only if it has exactly $q = \text{rank}(E)$ finite eigenvalues. The following lemma of [12] gives a useful characterization of regular, index one pencils.

Lemma 1.2. *The following statements are equivalent:*

1. *The pencil $\alpha E - \beta A$ is regular and has index less than or equal to one.*
2. $\text{rank} \left(\begin{bmatrix} E \\ T_\infty^T(E)A \end{bmatrix} \right) = \text{rank} (E + T_\infty(E)T_\infty^T(E)A) = n.$
3. $\text{rank} ([E, AS_\infty(E)]) = \text{rank} (E + AS_\infty(E)S_\infty^T(E)) = n.$
4. $T_\infty(E)^T AS_\infty(E)$ is nonsingular.
5. *If*

$$U^T E V = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$$

is the singular value decomposition (SVD) of E (with orthogonal matrices U, V and a nonsingular, diagonal matrix $\Sigma_r \in \mathbb{R}^{r \times r}$), then the $(n-r) \times (n-r)$ bottom right matrix A_{22} of $U^T A V$ is nonsingular.

In the notation of (3)–(4), classical solutions of (1a) take the form

$$x(t) = X_r z_1(t) + X_\infty z_2(t),$$

where

$$\begin{aligned} \dot{z}_1 &= J z_1 + Y_r^T B u \\ N \dot{z}_2 &= z_2 + Y_\infty^T B u \end{aligned}$$

and one has the explicit solution

$$\begin{aligned} z_1(t) &= e^{tJ} z_1(0) + \int_0^t e^{(t-s)J} Y_r^T B u(s) ds, \\ z_2(t) &= - \sum_{i=0}^{\nu-1} \frac{d^i}{dt^i} (N^i Y_\infty^T B u(t)). \end{aligned} \quad (5)$$

Equation (5) shows that the input functions must belong to some suitable function space \mathcal{U}_{ad} and, to ensure a smooth response for every continuous input $u(t)$, it is necessary for the system to be regular and have index less than or equal to one. Moreover, the possible values of the initial condition $x(0)$ are restricted. The initial state must be a member of the set of *consistent* initial conditions, i. e.,

$$\mathcal{S} \equiv \left\{ X_r z_1 + X_\infty z_2 \mid z_1 \in \mathbb{R}^r, z_2 = - \sum_{i=0}^{\nu-1} \left(\frac{d^i}{dt^i} (N^i Y_\infty^T B u)(0) \right), u(t) \in \mathcal{U}_{ad} \right\}.$$

The set of *reachable* states of (1a) from the *solution space* set \mathcal{S} of consistent initial conditions is \mathcal{S} itself.

2 Realization

In [16] a method for the *generalized realization problem* was presented. From given interpolation data, obtained by measurements from a real system or numerical simulation via a mathematical model, it generates a *descriptor system* of the form (1), i. e.,

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t). \end{aligned} \tag{6}$$

The generalized realization problem of [16] deals mainly with two cases.

1. In the scalar interpolation case, the given data consist of a vector of interpolation points $s = [s_i] \in \mathbb{C}^N$ and a vector of interpolation values $f = [f_i] \in \mathbb{C}^N$, and the realization problem constructs a transfer function of the form (2) satisfying the interpolation conditions

$$H(s_i) = C(s_i E - A)^{-1} B + D = f_i \quad i = 1, \dots, N.$$

2. In the matrix interpolation case, the interpolation points are again contained in a vector $s = [s_i] \in \mathbb{C}^N$. However, the interpolation values are summarized in form of a block matrix $F = [F_i]$ of N matrices $F_i \in \mathbb{C}^{p \times m}$ and the interpolation problem takes the following form: First, *right and left tangential data* are sampled by multiplying the matrix data F_i from the right (left) with arbitrary right (left) tangential directions such that as *right tangential interpolation conditions* we get

$$H(\lambda_i) r_i = (C(\lambda_i E - A)^{-1} B + D) r_i = w_i, \quad i = 1, \dots, \rho, \tag{7}$$

and as *left tangential interpolation conditions* we get

$$l_j H(\mu_j) = l_j (C(\mu_j E - A)^{-1} B + D) = v_j, \quad j = 1, \dots, \nu, \tag{8}$$

where r_i (l_j) are the right (left) tangential directions, w_i (v_j) are the right (left) tangential values, and λ_i (μ_j) are the right (left) interpolation points which are a subset of $\{s_1, \dots, s_N\}$.

The interpolation technique is realized in the `Matlab` codes `realization` and `loewner_mod` where the latter one is (a slightly modified version of) an m-File provided by the authors of [16]. Analytically, it can be shown that the obtained realization (6) is regular and *minimal*

(and thus *controllable and observable*), see [16]. However, there are no results regarding the index of the obtained descriptor system. Moreover, when the realization is computed numerically, the analytically guaranteed properties of regularity and minimality may be lost due to finite precision arithmetic. Thus, in general the realization obtained by computation may be non-regular, have index larger than one and miss certain controllability and observability properties, and therefore requires a regularization procedure which is described in the next section.

3 Controllability and Observability Conditions

Given the descriptor system (6), one or more of the following conditions are essential for most classical design aims, see e.g. [3, 5, 9].

$$\begin{aligned} \mathbf{C0}: \quad & \text{rank}[\alpha E - \beta A, B] = n \text{ for all } (\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}. \\ \mathbf{C1}: \quad & \text{rank}[\lambda E - A, B] = n \text{ for all } \lambda \in \mathbb{C}. \\ \mathbf{C2}: \quad & \text{rank}[E, AS_\infty(E), B] = n. \end{aligned} \tag{9}$$

A regular system is *completely controllable* or *C-controllable* if **C0** holds and is *strongly controllable* or *S-controllable* if **C1** and **C2** hold [5]. Complete controllability ensures that for any given initial and final states x_0, x_f there exists an admissible control that transfers the system from x_0 to x_f in finite time, while strong controllability ensures the same for any given initial and final states $x_0, x_f \in \mathcal{S}$ (the solution space).

Regular systems that satisfy condition **C2** are called *controllable at infinity* or *impulse controllable* [9]. For these systems, impulsive modes can be excluded by a suitable linear feedback.

Observability for descriptor systems is the dual of controllability. We define the following conditions:

$$\begin{aligned} \mathbf{O0}: \quad & \text{rank} \begin{bmatrix} \alpha E - \beta A \\ C \end{bmatrix} = n \text{ for all } (\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}. \\ \mathbf{O1}: \quad & \text{rank} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = n \text{ for all } \lambda \in \mathbb{C}. \\ \mathbf{O2}: \quad & \text{rank} \begin{bmatrix} E \\ T_\infty^T(E)A \\ C \end{bmatrix} = n. \end{aligned} \tag{10}$$

It is immediate that condition **O0** implies **O1** and **O2**. Moreover, **O1** and

$$\text{rank} \begin{bmatrix} E \\ C \end{bmatrix} = n, \tag{11}$$

together hold if and only if **O0** holds. A regular descriptor system is called *completely observable* or *C-observable* if condition **O0** holds and is called *strongly observable* or *S-observable* if conditions **O1** and **O2** hold. A regular system that satisfies condition **O2** is called *observable at infinity* or *impulse-observable*.

Conditions (9)–(11) are preserved under non-singular equivalence transformations as well as under state and output feedback, i. e., if the system satisfies **C0**, **C1**, or **C2**, then for any

non-singular $U \in \mathbb{R}^{n,n}$, $V \in \mathbb{R}^{n,n}$, $W \in \mathbb{R}^{m,m}$ and for any $F_1 \in \mathbb{R}^{m,n}$ and $F_2 \in \mathbb{R}^{m,p}$, the system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C})$, where

$$\tilde{E} = UEV, \quad \tilde{A} = UAV, \quad \tilde{B} = UBW \quad (12)$$

or

$$\tilde{E} = E, \quad \tilde{A} = A + BF_1, \quad \tilde{B} = B$$

or

$$\tilde{E} = E, \quad \tilde{A} = A + BF_2C, \quad \tilde{B} = B$$

also satisfies these conditions. Analogous properties hold for **O0**, **O1** and **O2**.

4 Regularization

In general, due to the finite precision arithmetic, it cannot be guaranteed that the system computed by the realization procedure presented in [16] satisfies the described regularity, controllability and observability conditions of Section 3. Therefore, it needs to be treated by a regularization procedure. The most general form of such a regularization procedure has been presented in [8]. It allows general non-square matrices E and A and it can be extended to general nonlinear systems. We briefly review this regularization procedure for the linear constant coefficient case. First, we write the state equation of system (6) in behavior form combining input and state to a joint vector $z = [x^T, u^T]^T$, i. e.,

$$\mathcal{E}\dot{z} = \mathcal{A}z \quad (13)$$

with $\mathcal{E} = [E, 0]$, $\mathcal{A} = [A, B]$ partitioned accordingly. Then following [7] we form a derivative array

$$\mathcal{M}_\ell \dot{z}_\ell = \mathcal{N}_\ell z_\ell, \quad (14)$$

where

$$\begin{aligned} (\mathcal{M}_\ell)_{i,j} &= \binom{i}{j} \mathcal{E}^{(i-j)} - \binom{i}{j+1} \mathcal{A}^{(i-j-1)}, \quad i, j = 0, \dots, \ell, \\ (\mathcal{N}_\ell)_{i,j} &= \begin{cases} \mathcal{A}^{(i)} & \text{for } i = 0, \dots, \ell, \quad j = 0, \\ 0 & \text{otherwise,} \end{cases} \\ (z_\ell)_j &= z^{(j)}, \quad j = 0, \dots, \ell. \end{aligned}$$

The subsequent Theorem follows from the more general results for variable coefficient systems, see [13]. It connects the derivative array with the strangeness index μ and is used for index reduction.

Theorem 4.1. *Consider system (13). There exists an integer μ such that the coefficients of the derivative array (14), (M_μ, N_μ) , associated with $(\mathcal{E}, \mathcal{A})$ have the following properties, where we set*

$$\hat{a} = a_\mu, \quad \hat{d} = d_\mu, \quad \hat{v} = v_0 + \dots + v_\mu. \quad (15)$$

1. $\text{rank } \mathcal{M}_\mu = (\mu + 1)n - \hat{a} - \hat{v}$, i. e., there exists a matrix Z of size $(\mu + 1)n \times (\hat{a} + \hat{v})$ and maximal rank satisfying $Z^T \mathcal{M}_\mu = 0$.
2. $\text{rank } Z^T \mathcal{N}_\mu [I_{n+m} \ 0 \ \dots \ 0]^T = \hat{a}$, i. e., Z can be partitioned as $Z = [Z_2 \ Z_3]$, with Z_2 of size $(\mu + 1)n \times \hat{a}$ and Z_3 of size $(\mu + 1)n \times \hat{v}$, such that $\hat{A}_2 = Z_2^T \mathcal{N}_\mu [I_{n+m} \ 0 \ \dots \ 0]^T$ has full row rank \hat{a} and $Z_3^T \mathcal{N}_\mu [I_{n+m} \ 0 \ \dots \ 0]^T = 0$. Furthermore, there exists a matrix T_2 of size $(n + m) \times (n + m - \hat{a})$ and maximal rank satisfying $\hat{A}_2 T_2 = 0$.

3. $\text{rank } \mathcal{E}(t)T_2 = \hat{d} = n - \hat{a} - v_\mu$, i. e., there exists a matrix Z_1 of size $n \times \hat{d}$ and maximal rank satisfying $\text{rank}(\hat{E}_1 T_2) = \hat{d}$ with $\hat{E}_1 = Z_1^T \mathcal{E}$.

Furthermore, system (13) has the same solution set as the system

$$\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix} \dot{z} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} z, \quad (16)$$

where $\hat{E}_1 = Z_1^T \mathcal{E}$, $\hat{A}_1 = Z_1^T \mathcal{A}$ and $\hat{A}_2 = Z_2^T \mathcal{N}_\mu [I_{n+m} \ 0 \ \cdots \ 0]^T$.

The smallest number μ for which Theorem 4.1 holds is called the *strangeness index*. The differential-algebraic system (16) is strangeness-free, i. e., its strangeness index is zero. Its coefficients can be computed by using three nullspace computations, which are carried out via SVDs or QR decompositions with column pivoting (cf. [11]) as long as this is feasible in the available computing environment. The system (16) is a *reformulation* of (13) (using the original model and its derivatives) without changing the solution set, since no transformation of the vector z has been made. The constructed submatrices \hat{A}_1 and \hat{A}_2 have been obtained from the block matrix

$$\begin{bmatrix} A & B \\ \dot{A} & \dot{B} \\ \vdots & \vdots \\ A^{(\mu)} & B^{(\mu)} \end{bmatrix}$$

by transformations from the left. This has two immediate consequences [14]. First, derivatives of the input function u are nowhere needed, i. e., although formally the derivatives of u occur in the derivative array, they do not occur in the form (16), and hence, we do not have any additional smoothness requirements for the input function u .

Second, it follows from the construction of \hat{A}_1 and \hat{A}_2 that the partitioning into the part stemming from the original states x and the original controls u is not mixed up. Including the output equation, we obtain a reformulated system of the form

$$E_1 \dot{x} = A_1 x + B_1 u, \quad (17a)$$

$$0 = A_2 x + B_2 u, \quad (17b)$$

$$0 = 0, \quad (17c)$$

$$y = Cx + Du, \quad (17d)$$

where

$$E_1 = \hat{E}_1 \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad A_i = \hat{A}_i \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad B_i = \hat{A}_i \begin{bmatrix} 0 \\ I_m \end{bmatrix}, \quad i = 1, 2.$$

Here E_1, A_1 have size $d \times n$, while E_2, A_2 are of size $a \times n$. The equations in (17c) can just be removed from the system and we continue with the modified model of $d + a$ equations

$$\begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix} \dot{z} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix} z$$

together with given initial conditions. Consistency of initial values can easily be checked, they have to satisfy the equation

$$A_2 x(t_0) + B_2 u(t_0) = 0,$$

which (if B_2 does not vanish) represents a restriction on the initial value of the control u .

In (17a) and (17b), we have $d + a$ equations and n variables in x and m variables in u . In order for this system to be *regular*, i. e., uniquely solvable for all sufficiently smooth inputs u , and all consistent initial conditions, we would need that $d + a = n$.

If $d + a < n$, then for given u we cannot expect a unique solution, i. e., the system is not regular and we can just attach $n - (d + a)$ variables from x to u and if $d + a > n$, then we just attach $d + a - n$ of the input variables in u to the vector x . There is freedom in the choice of the variables that are chosen for reinterpretation, and ideally the selection should be done in such a way that the resulting descriptor system is regular if the input $u = 0$ is used, but this is not necessary. Note that we must also change the output equation by moving appropriate columns from D to C or vice versa. As a result of the reinterpretation, we obtain a new system

$$\begin{aligned}\tilde{E}_1 \dot{\tilde{x}} &= \tilde{A}_1 \tilde{x} + \tilde{B}_1 \tilde{u}, \\ 0 &= \tilde{A}_2 \tilde{x} + \tilde{B}_2 \tilde{u}, \\ y &= \tilde{C} \tilde{x} + \tilde{D} \tilde{u},\end{aligned}$$

where now the matrices $\begin{bmatrix} \tilde{E}_1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} \tilde{A}_1 \\ \tilde{A}_2 \end{bmatrix}$ are square of size $\tilde{n} = d + a$, and $\begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{bmatrix}$ is of size $\tilde{n} \times \tilde{m}$ with $\tilde{m} = n + m - \tilde{n}$.

It is often also useful to remove the feed-through term $\tilde{D} \tilde{u}$ in the output equation. This can be done by expanding the state dimension by introducing $\tilde{x}_{aux} := \tilde{D} \tilde{u}$ and rewriting the system as

$$\begin{aligned}\bar{E}_1 \dot{\bar{x}} &= \bar{A}_1 \bar{x} + \bar{B}_1 \tilde{u}, \\ 0 &= \bar{A}_2 \bar{x} + \bar{B}_2 \tilde{u}, \\ y &= \bar{C} \bar{x},\end{aligned}$$

with

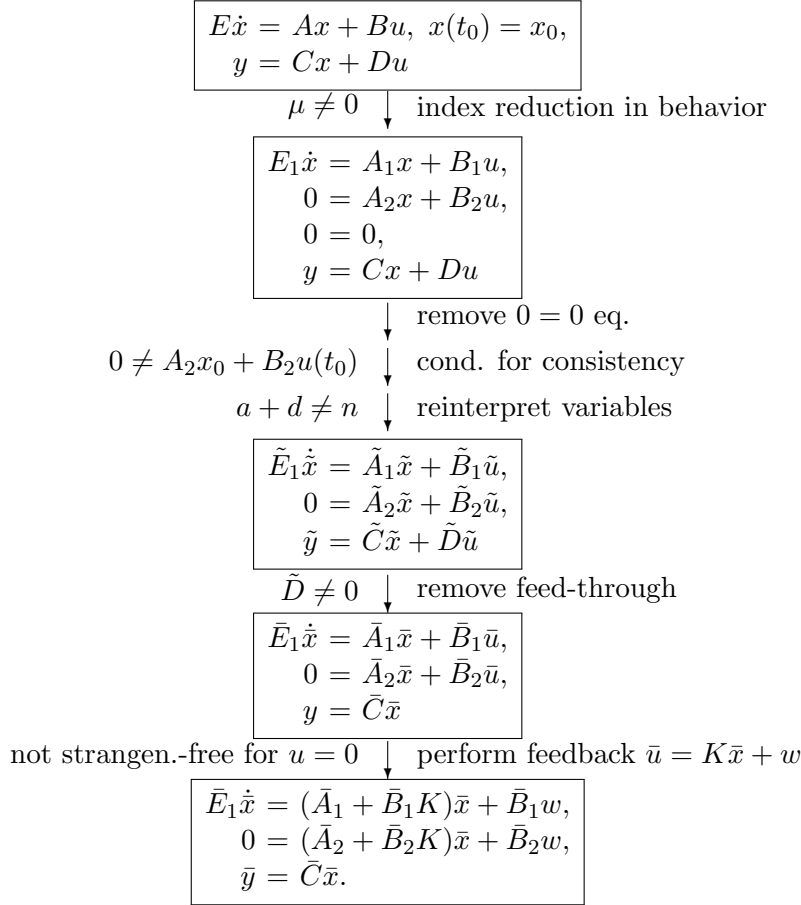
$$\begin{aligned}\bar{x} &= \begin{bmatrix} \tilde{x} \\ \tilde{x}_{aux} \end{bmatrix}, \bar{E}_1 = \begin{bmatrix} \tilde{E}_1 & 0 \end{bmatrix}, \bar{A}_1 = \begin{bmatrix} \tilde{A}_1 & 0 \end{bmatrix}, \bar{A}_2 = \begin{bmatrix} \tilde{A}_2 & 0 \\ 0 & I_p \end{bmatrix}, \bar{B}_1 = \tilde{B}_1, \\ \bar{B}_2 &= \begin{bmatrix} \tilde{B}_2 \\ -\tilde{D} \end{bmatrix}, \bar{C} = \begin{bmatrix} \tilde{C} & I_p \end{bmatrix}.\end{aligned}$$

This method of removing the feed-through term leads to an increase of the state dimension by p , i. e., from \tilde{n} to $\bar{n} = \tilde{n} + p$. The resulting system may again be of index higher than one as a free system with $\tilde{u} = 0$. But in this case, see [14], there exists a linear feedback $\tilde{u} = K \bar{x} + w$, with $K \in \mathbb{R}^{\tilde{m}, \bar{n}}$ such that in the closed loop system

$$\bar{E} \dot{\bar{x}} = (\bar{A} + \bar{B}K) \bar{x} + \bar{B}w, \quad \bar{x}(t_0) = \bar{x}_0, \quad (18a)$$

$$y = \bar{C} \bar{x}, \quad (18b)$$

the matrix function $(\bar{A}_2 + \bar{B}_2 K) \bar{T}_2'$ is nonsingular, and \bar{T}_2' is a matrix valued function that spans the kernel of \bar{E}_1 . This implies that the differential-algebraic equation system in (18a) is regular and of index at most one as a free system with $w = 0$, see Lemma 1.2. We summarize the whole regularization procedure in the following diagram, see [8].



In the following we assume that the system has been regularized to the form (18) and furthermore that \bar{B} and \bar{C}^T have full column rank. Otherwise, we can just reduce the input vector or the output vector. In abuse of notation, we denote the resulting system again in the original notation

$$\begin{aligned} E\dot{x} &= Ax + Bu, \quad x(t_0) = x_0, \\ y &= Cx, \end{aligned} \tag{19}$$

Note that the resulting system may not satisfy the desired controllability and observability conditions associated with the finite spectrum, and even if, then it may be close to a system that does not satisfy these conditions. To remove uncontrollable and unobservable finite parts, i. e., to make the system minimal, some further transformations may be necessary. In the following section we discuss condensed forms under orthogonal transformations which can be used to check all the controllability conditions from Section 3.

5 Condensed Forms

To verify the controllability and observability conditions, using equivalence transformations such as (12), the regularized system (19) is transformed to a condensed form that reveals these properties. The following condensed form has been presented in full generality in [5]. It uses only real orthogonal transformations and can be computed using algorithms that are

numerically stable in the sense that in finite precision arithmetic, the computed condensed form is what would have been obtained using exact arithmetic from a rounding-error-small perturbation of the original descriptor system. In the following we adopt the notation that a matrix Σ_j is a non-singular j -by- j diagonal matrix, and 0 denotes the null-matrix of any size.

Unfortunately, all condensed forms rely on numerical rank decisions of transformed submatrices of E , A , B and C . This is a serious problem, since arbitrarily small perturbations of a rank deficient matrix may change its rank.

Theorem 5.1. [6] *Let $E, A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,m}$, and $C \in \mathbb{R}^{p,n}$, where B and C are of full column and row rank, respectively. Then, there exist orthogonal matrices $U, V \in \mathbb{R}^{n,n}$, $W \in \mathbb{R}^{m,m}$, and $Y \in \mathbb{R}^{p,p}$ such that*

$$U^T E V = \begin{matrix} & t_1 & n-t_1 \\ t_1 & \Sigma_{t_1} & 0 \\ n-t_1 & 0 & 0 \end{matrix}, \quad (20a)$$

$$U^T B W = \begin{matrix} & k_1 & k_2 \\ t_1 & B_{11} & B_{12} \\ t_2 & B_{21} & 0 \\ t_3 & B_{31} & 0 \\ n-t_1-t_2-t_3 & 0 & 0 \end{matrix}, \quad (20b)$$

$$Y^T C V = \begin{matrix} & t_1 & s_2 & t_5 & n-t_1-s_2-t_5 \\ \ell_1 & C_{11} & C_{12} & C_{13} & 0 \\ \ell_2 & C_{21} & 0 & 0 & 0 \end{matrix}, \quad (20c)$$

$$U^T A V = \begin{matrix} & t_1 & s_2 & t_5 & t_4 & t_3 & s_6 \\ t_1 & A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} \\ t_2 & A_{21} & A_{22} & A_{23} & A_{24} & 0 & 0 \\ t_3 & A_{31} & A_{32} & A_{33} & A_{34} & \Sigma_{t_3} & 0 \\ t_4 & A_{41} & A_{42} & A_{43} & \Sigma_{t_4} & 0 & 0 \\ t_5 & A_{51} & 0 & \Sigma_{t_5} & 0 & 0 & 0 \\ t_6 & A_{61} & 0 & 0 & 0 & 0 & 0 \end{matrix}. \quad (20d)$$

The matrix B_{12} has full column rank, C_{21} has full row rank, and the matrices

$$\begin{bmatrix} B_{21} \\ B_{31} \end{bmatrix} \in \mathbb{C}^{k_1 \times k_1}, \quad \begin{bmatrix} C_{12} & C_{13} \end{bmatrix} \in \mathbb{C}^{\ell_1 \times \ell_1}$$

are square and non-singular and are of dimension $k_1 = t_2 + t_3$ and $\ell_1 = s_2 + t_5$, respectively. Here t_j , s_j , k_j and ℓ_j are non-negative integers displaying the number of rows or columns in the corresponding block row or column of the matrices. A zero value of one of these integers indicates that the corresponding block row or column does not appear.

As a corollary we can characterize controllability and observability conditions of Section 3.

Corollary 5.2. *Consider a system of the form (19) and let the system be transformed to the condensed form (20a)–(20d) of Theorem 5.1.*

1. *The pair (E, A) is regular and of index at most one if and only if $s_6 = t_6 = 0$ and A_{22} is nonsingular.*

2. Condition **C2** holds if and only if $t_6 = 0$.
3. Condition **O2** holds if and only if $s_6 = 0$.
4. $\text{rank}[E, B] = t_1 + t_2 + t_3$, and thus $\text{rank}[E, B] = n$ if and only if $t_4 = t_5 = t_6 = 0$.
5. $\text{rank} \begin{bmatrix} E \\ C \end{bmatrix} = t_1 + s_2 + t_5$, and thus $\text{rank} \begin{bmatrix} E \\ C \end{bmatrix} = n$ if and only if $t_4 = t_3 = s_6 = 0$.
6. $\text{rank} \begin{bmatrix} E & B \\ C & 0 \end{bmatrix} = t_1 + t_2 + s_2 + t_3 + t_5 + \min(\ell_2, k_2)$.

If we have computed the condensed form from the regularized system (19) then we should have $t_6 = s_6 = 0$ and the system is of index at most one as a free system. The staircase form allows to check whether the regularization procedure has been successful.

6 Matlab Functions in detail

In the following several **Matlab** functions are presented, which create a regularized Loewner realization based on tangential interpolation data of the transfer function. It strongly builds on the procedure of [16], see Section 2, followed by a regularization based on the methods and results outlined in the previous sections. As a result, we obtain a realization which is regular and strangeness-free as well as completely controllable and observable.

6.1 Realization

Syntax

```
[E,A,B,C,D,mu,la,V,W,L,R] = realization(S,F)
[E,A,B,C,D,mu,la,V,W,L,R,U_trans,V_trans,W_trans,Y_trans,...
 L_trans,R_trans,Feedb] = realization(S,F)
[E,A,B,C,D,mu,la,V,W,L,R,U_trans,V_trans,W_trans,Y_trans,...
 L_trans,R_trans,Feedb] = realization(S,F,tol)
[E,A,B,C,D,mu,la,V,W,L,R,U_trans,V_trans,W_trans,Y_trans,...
 L_trans,R_trans,Feedb] = realization(S,F,tol,sindexflag)
```

Arguments

The following table lists the input arguments of the function **realization**.

S	Vector of length N of interpolation points (is split into two disjoint interpolation point sets mu and lambda)
F	Array which contains the transfer function values at points S ; this is either a vector of length N (scalar interpolation case) or a $p \times m \times N$ array consisting of N $p \times m$ matrices (matrix interpolation case)
tol	scalar specifying the tolerance value for rank decisions in function hypo (default: 10*eps , where eps is the floating-point relative accuracy 2^{-52})
sindexflag	boolean (default: true); if true , index reduction and regularization are performed; if false , Loewner realization is provided without post-processing steps

The following table lists the output arguments of the function `realization`.

<code>[E,A,B,C,D]</code>	matrices corresponding to a system which is regularized and strangeness-free (if <code>sindexflag</code> is set to <code>true</code>) and whose transfer function $H(z) = C(zE - A)^{-1}B + D$ interpolates the data (S,F)
<code>mu</code>	vector of size ν containing the left interpolation points μ_j (cf. (8))
<code>la</code>	vector of size ρ containing the right interpolation points λ_i (cf. (7))
<code>V</code>	<i>scalar interpolation case</i> : vector of size ν containing the left interpolation values v_j belonging to μ_j (cf. (8) with $l_j = 1$) <i>matrix interpolation case</i> : matrix of dimension $\nu \times m$ containing the left interpolation values $v_j \in \mathbb{C}^{1,m}$ (as rows) generated by random left tangential directions $l_j \in \mathbb{C}^{1,p}$, $j = 1, \dots, \nu$ (cf. (8))
<code>W</code>	<i>scalar interpolation case</i> : vector of size ρ containing the right interpolation values w_i belonging to λ_i (cf. (7) with $r_i = 1$) <i>matrix interpolation case</i> : matrix of dimension $p \times \rho$ containing the right interpolation values $w_i \in \mathbb{C}^p$ (as columns) generated by random right tangential directions $r_i \in \mathbb{C}^m$, $i = 1, \dots, \rho$ (cf. (7))
<code>L</code>	matrix of dimension $\nu \times p$ containing (random) left tangential directions $l_j \in \mathbb{C}^{1,p}$ as rows (set to one in scalar interpolation case)
<code>R</code>	matrix of dimension $m \times \rho$ containing (random) right tangential directions $r_i \in \mathbb{C}^m$ as columns (set to one in scalar interpolation case)
<code>U_trans, V_trans, W_trans, Y_trans</code>	matrices corresponding to the transformation matrices U , V , W , and Y of Theorem 5.1 (see also description of function <code>staircase</code>)
<code>L_trans, R_trans</code>	transformation matrices from block Gaussian elimination in function <code>regularization</code>
<code>Feedb</code>	feedback matrix from function <code>regularization</code> to make (E, A) regular

Remark 6.1. The vectors la and mu together form S such that $\nu + \rho = N$ and the sizes of la and mu differ by one when N is odd and are the same when N is even. Furthermore, if S contains values with non-zero imaginary part the complex conjugate values are added to S to ensure that the output realization `[E,A,B,C,D]` consists of real-valued matrices.

Description

`[E,A,B,C,D,mu,la,V,W,L,R] = realization(S,F)` constructs matrices E , A , B , C , and D such that the transfer function $H(s) = C(sE - A)^{-1}B + D$ interpolates the given data as described in Section 2. First, the Loewner matrix \mathbb{L} and the shifted Loewner matrix \mathbb{L}_σ as well as the corresponding matrices V, W, L, R and vectors μ, λ are constructed by use of the Matlab function `[LL,sLL,mu,la,V,W,L,R] = loewner_mod(S,F)`. Then two cases have to be considered. If $\rho = \nu$ and $\det(\tilde{s}\mathbb{L} - \mathbb{L}_\sigma) \neq 0$ for all $\tilde{s} \in \{\lambda_i\} \cup \{\mu_j\}$, which implies that $\tilde{s}\mathbb{L} - \mathbb{L}_\sigma$ is quadratic and nonsingular (regular case), the Loewner realization (E, A, B, C) is given by $E = -\mathbb{L}$, $A = -\mathbb{L}_\sigma$, $B = V$, and $C = W$ resulting in a system with the desired interpolation properties. The second case is the nonregular case. To make sure that we still

get a regular system, we need to ensure that

$$\text{rank}(\tilde{s}\mathbb{L} - \mathbb{L}_\sigma) = \text{rank} \begin{bmatrix} \mathbb{L} & \mathbb{L}_\sigma \end{bmatrix} = \text{rank} \begin{bmatrix} \mathbb{L} \\ \mathbb{L}_\sigma \end{bmatrix}, \quad \text{for all } \tilde{s} \in \{\lambda_i\} \cup \{\mu_j\}. \quad (21)$$

If this condition is satisfied, we choose an arbitrary $\tilde{s} \in \{\lambda_i\} \cup \{\mu_j\}$ and compute the skinny SVD $(\tilde{s}\mathbb{L} - \mathbb{L}_\sigma) = Y\Sigma X^T$, see [11] with a nonsingular diagonal matrix Σ and transformation matrices Y and X with pairwise orthonormal columns. In this case the Loewner realization, see [16], is given by

$$E = -Y^T \mathbb{L} X, \quad A = -Y^T \mathbb{L}_\sigma X, \quad B = Y^T V, \quad C = W X, \quad D = 0.$$

Even if the regularity of the matrix pencil (E, A) is guaranteed analytically, in the finite precision case, we cannot be sure about this. Further important properties as the index, controllability, and observability are also unknown, in general. Thus, to obtain a regular, strangeness-free, completely observable, and completely controllable system, some further steps have to be performed (only executed if `sindexflag` is true).

First, the index is reduced by applying Theorem 4.1 using the function `hypo`. The resulting system is strangeness-free and is of the form (17). The vanishing equations can be neglected such that the number of equations decreases to $a + d$. If this number differs from the number of variables, either some of the components of x have to be attached to the vector u or vice versa. This changes the input dimension m such that the size of the transfer function does not fit to the tangential interpolation data anymore. However, if the problem is well-posed, this case should not occur.

To obtain more insight into the controllability and observability properties of the realization, the function `staircase` is called, which computes the condensed form of the realization (E, A, B, C) according to Theorem 5.1. The system matrices corresponding to the condensed form are denoted with EC , AC , BC , and CC , respectively. If B does not have full column rank or if C does not have full row rank, the resulting zero columns in BC or zero rows in CC are canceled which decreases the dimensions of L , R , V , and W accordingly.

The subsequent regularization procedure, performed by the function `regularization`, eliminates the non-controllable and non-observable parts leading to system matrices E_{reg} , A_{reg} , B_{reg} , and C_{reg} as

$$E_{reg} = \begin{bmatrix} E_{11} & \\ & 0 \end{bmatrix}, \quad A_{reg} = \begin{bmatrix} A_{11} & \\ & A_{22} \end{bmatrix},$$

$$B_{reg} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix}, \quad C_{reg} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & 0 \end{bmatrix},$$

where the blocks E_{11} and A_{22} are nonsingular. Consequently, the pencil $(\tilde{s}E_{reg} - A_{reg})$ is regular. Finally, expanding the product of the matrices, one gets

$$\begin{bmatrix} C_{11} \\ C_{21} \end{bmatrix} (sE_{11} - A_{11})^{-1} \begin{bmatrix} B_{11} & B_{12} \end{bmatrix} - C_{12} A_{22}^{-1} B_{21},$$

such that we can decrease the state-space dimension of the realization by setting

$$E = E_{11}, \quad A = A_{11}, \quad B = \begin{bmatrix} B_{11} & B_{12} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} \\ C_{21} \end{bmatrix}, \quad D = -C_{12} A_{22}^{-1} B_{21}.$$

6.2 Hypo

Syntax

`[E1_hat,A1_hat,A2_hat,d,a,v,mu_max,sig] = hypo(E,A,mu,tol,varargin)`

Arguments

The following table lists the input arguments of the function `hypo`.

<code>E</code>	matrix \mathcal{E} of the system's behavior form as in (13), i. e., $\mathcal{E} = [E, 0]$
<code>A</code>	matrix \mathcal{A} of the system's behavior form as in (13), i. e., $\mathcal{A} = [A, B]$
<code>mu</code>	corresponds to the index ℓ of the inflated system (14) (default: 0)
<code>tol</code>	scalar specifying the tolerance value for rank decisions (default: <code>10*eps</code>)
<code>varargin</code>	contains $v_0, \dots, v_{\ell-1}$ (default: empty)

The following table lists the output arguments of the function `hypo`.

<code>E1_hat, A1_hat,</code>	
<code>A2_hat</code>	blocks of the reformulated system (16)
<code>d</code>	number of differential equations (\hat{d})
<code>a</code>	number of algebraic equations (\hat{a})
<code>v</code>	number of vanishing equations (v_μ)
<code>mu_max</code>	strangeness index of the original system (\mathcal{E}, \mathcal{A})
<code>sig</code>	error resulting from rank decision

Description

The function `hypo` successively inflates the system $\mathcal{E}\dot{z} = \mathcal{A}z$ by differentiation which leads to inflated systems $\mathcal{M}_\ell \dot{z}_\ell = \mathcal{N}_\ell z_\ell$ with ℓ starting at 0 and being incremented by one in each step. This procedure is continued until the rank conditions of Theorem 4.1 are fulfilled yielding μ , \hat{d} , \hat{a} , and \hat{v} . The matrix Z is computed by means of an SVD of \mathcal{M}_ℓ using those left singular vectors that lie in the left null space of \mathcal{M}_ℓ . The first singular value that is considered to be negligibly small (during rank decision based on `tol`) is used as an error measurement of the procedure.

The matrices Z_2 and T_2 are determined based on the SVD

$$Z^T \mathcal{N}_\ell \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix}^T = USV^T.$$

Z_2 consists of the first \hat{a} columns of ZU , i. e., such that $Z_2^T \mathcal{N}_\ell \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix}^T$ has full row rank and T_2 consists of those columns of V lying in the right null space of the matrix $Z_2^T \mathcal{N}_\ell \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix}^T$. The difference between the number of columns of Z and the number of algebraic constraints \hat{a} is equal to \hat{v} , cf. Theorem 4.1.

Finally, Z_1 is determined by calculating a QR-decomposition of $\mathcal{E}T_2$ and by choosing \hat{d} columns of Q such that $Z_1^T \mathcal{E}T_2$ has full rank \hat{d} . If the sum $\hat{d} + \hat{a} + v_\ell$ (using $v_\ell = \hat{v} - \sum_{i=0}^{\ell-1} v_i$, cf. (15)) differs from the number of equations of the system $\mathcal{E}\dot{z} = \mathcal{A}z$, the index ℓ is increased by one and `hypo` is called with `varargin` containing $v_0, \dots, v_{\ell-1}$. Otherwise the index reduction is complete and we set $\hat{E}_1 = Z_1^T \mathcal{E}$, $\hat{A}_1 = Z_1^T \mathcal{A}$ and $\hat{A}_2 = Z_2^T \mathcal{N}_\ell \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix}^T$. The number μ_{max} corresponds to the smallest index ℓ needed to satisfy Theorem 4.1. This number is equal to the strangeness index of the original system $(\mathcal{E}, \mathcal{A})$.

6.3 Staircase

Syntax

`[EC,AC,BC,CC,U,V,W,Y,t,s,k,l] = staircase(E,A,B,C)`

Arguments

The following table lists the input arguments of the function `staircase`.

E	$n1 \times n2$ matrix
A	$n1 \times n2$ matrix
B	$n1 \times m$ matrix
C	$p \times n2$ matrix

The following table lists the output arguments of the function `staircase`.

EC, AC, BC, CC	condensed form of the input system matrices (E, A, B, C) according to Theorem 5.1
U, V, W, Y	orthogonal matrices that transform E, A, B , and C to condensed form, i. e., $EC = U^T E V$, $AC = U^T A V$, $BC = U^T B W$, and $CC = Y^T C V$
t, s, k, l	vectors containing the block dimensions of the condensed form, see Theorem 5.1

Description

The algorithm follows the constructive proof of Theorem 5.1, which is presented in [6]. For that, numerous SVDs are used to transform the input matrices E, A, B , and C into the form

$$\begin{aligned}
 EC &= \begin{bmatrix} \Sigma_E & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, & AC &= \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} \\ A_{21} & A_{22} & A_{23} & A_{24} & 0 & 0 \\ A_{31} & A_{32} & A_{33} & A_{34} & \Sigma_{t_3} & 0 \\ A_{41} & A_{42} & A_{43} & \Sigma_{t_4} & 0 & 0 \\ A_{51} & 0 & \Sigma_{t_5} & 0 & 0 & 0 \\ A_{61} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \\
 BC &= \begin{bmatrix} B_{11} & B_{12} & 0 \\ B_{21} & 0 & 0 \\ B_{31} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, & CC &= \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & 0 & 0 \\ C_{21} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},
 \end{aligned} \tag{22}$$

where **EC** and **AC** are of size $(t_1 + t_2 + t_3 + t_4 + t_5 + t_6) \times (t_1 + s_2 + t_5 + t_4 + t_3 + s_6)$, **BC** is of size $(t_1 + t_2 + t_3 + t_4 + t_5 + t_6) \times (k_1 + k_2 + (m - k_1 - k_2))$ and **CC** is of size $(l_1 + l_2 + (p - l_1 - l_2)) \times (t_1 + s_2 + t_5 + t_4 + t_3 + s_6)$. Accordingly, we have $n1 = t_1 + t_2 + t_3 + t_4 + t_5 + t_6$ and $n2 = t_1 + s_2 + t_5 + t_4 + t_3 + s_6$. Note that the difference between (22) and the condensed form presented in Theorem 5.1 is that (22) allows for general input matrices B and C without assuming full row or column rank. During the algorithm also the transformation matrices are built such that

$$\mathbf{EC} = U^T \mathbf{E} V, \quad \mathbf{AC} = U^T \mathbf{A} V, \quad \mathbf{BC} = U^T \mathbf{B} W, \quad \mathbf{CC} = Y^T \mathbf{C} V,$$

with $U \in \mathbb{C}^{n1 \times n1}$, $V \in \mathbb{C}^{n2 \times n2}$, $W \in \mathbb{C}^{m \times m}$, and $Y \in \mathbb{C}^{p \times p}$.

6.4 Regularization

Syntax

`[E,A,B,C,L_trans,R_trans,Feedb]= regularization(EC,AC,BC,CC,t,s,k,l)`

Arguments

The following table lists the input arguments of the function `regularization`.

<code>EC, AC, BC, CC</code>	matrices in condensed form generated by the function <code>staircase</code>
<code>t, s, k, l</code>	vectors containing the block dimensions of the condensed form generated by the function <code>staircase</code>

The following table lists the output arguments of the function `regularization`.

<code>E, A, B, C</code>	controllable and observable system where (E, A) is regular
<code>L_trans, R_trans</code>	left and right transformation matrices such that $E = L_{trans} \mathbf{E} C R_{trans}$, $A = L_{trans} \mathbf{A} C R_{trans} + L_{trans} \mathbf{B} C \mathbf{F} eedb$, $B = L_{trans} \mathbf{B} C$ and $C = \mathbf{C} C R_{trans}$
<code>Feedb</code>	feedback matrix, which ensures that the block A_{22} of A is nonsingular

Description

In the function `regularization`, first it is checked whether the input system can be made regular and of index one. This means that we have to ensure, that the matrices \mathbf{EC} and \mathbf{AC} are quadratic and that $t_6 = s_6 = 0$. If this is true, the matrices have the following form:

$$\mathbf{EC} = \begin{bmatrix} \Sigma_E & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{AC} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ A_{21} & A_{22} & A_{23} & A_{24} & 0 \\ A_{31} & A_{32} & A_{33} & A_{34} & \Sigma_{t_3} \\ A_{41} & A_{42} & A_{43} & \Sigma_{t_4} & 0 \\ A_{51} & 0 & \Sigma_{t_5} & 0 & 0 \end{bmatrix},$$

$$\mathbf{BC} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \\ B_{31} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{CC} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & 0 \\ C_{21} & 0 & 0 & 0 & 0 \end{bmatrix},$$

where the block A_{22} is quadratic ($s_2 \stackrel{!}{=} t_2$) and the zero columns and rows of \mathbf{BC} and \mathbf{CC} are canceled out in the function `realization`.

The blocks Σ_{t_i} in **AC** are invertible diagonal matrices such that a block Gaussian elimination can be performed to eliminate the corresponding rows and columns inside **AC** leading to

$$\mathbf{A1} = \begin{bmatrix} A_{11} & A_{12} & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \Sigma_{t_3} \\ 0 & 0 & 0 & \Sigma_{t_4} & 0 \\ 0 & 0 & \Sigma_{t_5} & 0 & 0 \end{bmatrix}.$$

BC and **CC** are transformed accordingly to **B1** and **C1** without changing the block structure while **EC** stays completely unchanged due to its zero-block structure. In the end we are only interested in the system's transfer function $H(s) = C(sE - A)^{-1}B + D$. Thus, we can restrict ourselves to the upper left 2×2 block of **EC** and **A1**, since by multiplying the lower right part of $(sE - A)^{-1}$, namely the block

$$\begin{bmatrix} 0 & 0 & -\Sigma_{t_5}^{-1} \\ 0 & -\Sigma_{t_4}^{-1} & 0 \\ -\Sigma_{t_3}^{-1} & 0 & 0 \end{bmatrix},$$

by the corresponding blocks of **B1** and **C1**, it cancels out and, hence, it does not contribute to the transfer function. The system can be reduced to

$$\begin{aligned} E_{new} &= \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, & A_{new} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \\ B_{new} &= \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix}, & C_{new} &= \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & 0 \end{bmatrix}, \end{aligned} \tag{23}$$

where, by abuse of notation, we have redefined the naming of the matrix blocks, i. e., A_{11} in (23) is not necessarily the same as A_{11} in (22) and so on.

If the block A_{22} is singular, then the pencil $(sE_{new} - A_{new})$ will not be strangeness-free. In this case a feedback is added using the fact that the block B_{21} is invertible by its construction in **staircase**. We construct a matrix *Feedb* such that

$$A_{new} + B_{new}Feedb = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \sigma I \end{bmatrix} = A_{new2},$$

where I is the identity matrix and σ denotes the smallest singular value of the first block row of A_{new} . Using block Gaussian elimination we can then transform A_{new2} into block diagonal form and obtain the desired regularized system together with the transformation matrices L_{trans} and R_{trans} .

7 Numerical Example

In this section the Loewner framework, endowed with the index reduction and regularization procedure outlined in Section 6, is illustrated by means of an example from the Oberwolfach Model Reduction Benchmark Collection [2]. We consider the nonlinear heat transfer in a one-dimensional beam discussed in [15]. A schematic illustration of the system is depicted in Figure 1. For the sake of simplicity we restrict ourselves to the single-input single-output (SISO) case in contrast to the multiple-input multiple-output (MIMO) system considered in [15].

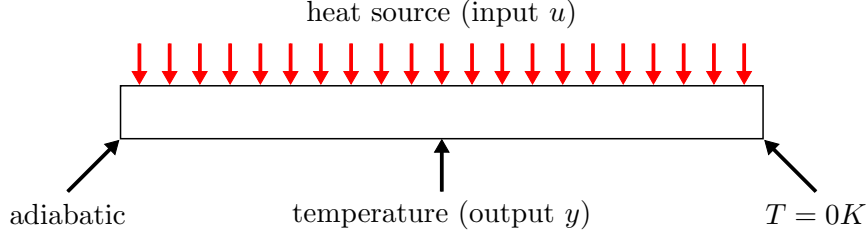


Figure 1: 1D-Beam with heat source (input) and measured temperature (output)

The governing equation of the physical system is a parabolic partial differential equation describing the temporal progress of the spatial temperature distribution along the beam. However, instead of the absolute temperature T_{abs} , a relative temperature T is considered, i.e., $T = T_{abs} - T_{ref}$ with respect to a reference temperature T_{ref} . The initial condition is chosen homogeneously as $T = 0$ over the whole beam at time $t_0 = 0$. Furthermore, at the left boundary an adiabatic end is assumed, i.e., zero temperature gradient, and at the right boundary the relative temperature is equal to zero for all times $t > 0$ [15].

In this example we are rather interested in the input-output (I/O) behavior than in the time progress of the entire temperature distribution. As an input a heat source is applied affecting the whole beam homogeneously and the temperature at the middle of the beam represents the system output.

Moreover, a nonlinearity comes into play by considering a thermal conductivity which depends on the temperature polynomially, i.e.,

$$k(T) = \sum_{i=0}^N a_i T^i$$

with given coefficients a_i . After modeling, discretization and renaming of variables ($T \rightarrow x$) one obtains a dynamical system of the form

$$\begin{aligned} E\dot{x} &= Ax + bu + f(x), \\ y &= c^T x, \end{aligned} \tag{24}$$

where $x \in \mathbb{R}^n$ denotes the state vector (discrete approximation of temperature), u the input (heat source) and y the output (temperature at the middle of the beam). Furthermore, $E, A \in \mathbb{R}^{n,n}$ and $b, c \in \mathbb{R}^n$ represent the linear part and the function $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ constitutes the nonlinear part of the dynamical system. More details regarding the modeling and discretization may be found in [15].

Depending on the mesh size, there are two systems of different dimensions available within the Oberwolfach Model Reduction Benchmark Collection: $n = 15$ and $n = 410$. Since the main intention is to illustrate the need for the regularized Loewner approach, we choose the system of dimension $n = 15$ due to the significantly smaller simulation times.

In order to use the Loewner method we need sampled data of the transfer function of the system. Since a nonlinear dynamical system is considered, there is only little hope to find an analytic expression for the transfer function of the system. The idea is instead to utilize the system's impulse response and determine a linear transfer function describing the input-output behavior of the system for the chosen input. Due to the nonlinearity of the system

the obtained transfer function has only a limited validity range with its size depending on the impact of the nonlinearity on the I/O map.

Since an actual impulse response is numerically unfeasible, instead we create the step response and differentiate it numerically, in order to obtain an approximation of the impulse response, as in [4]. The discrete values of the impulse response are equal to the Markov parameters h_k ($k = 0, 1, \dots$) of the corresponding discrete-time system leading to the discrete-time transfer function

$$\tilde{H}(z) = \sum_{k=0}^{\infty} h_k z^{-k}. \quad (25)$$

Since the impulse response of the considered system approaches zero for large time values, the same holds for the Markov parameters with high index. Consequently, the infinite sum of equation (25) may be truncated while retaining a reasonable level of accuracy. For applying the Loewner approach, the transfer function is expected to map from the Laplace transforms of the inputs to the Laplace transforms of the outputs, cf. Section 1. However, the obtained transfer function \tilde{H} refers to the Z-domain. In order to obtain an expression for the transfer function of the continuous-time system the bilinear transformation is used to transform from the Z-domain to the Laplace domain [17], i. e.,

$$z = \frac{1 + \frac{\Delta t}{2}s}{1 - \frac{\Delta t}{2}s}$$

leading to

$$H(s) = \tilde{H}\left(\frac{1 + \frac{\Delta t}{2}s}{1 - \frac{\Delta t}{2}s}\right),$$

where Δt denotes the sampling time interval, which is equal to the time step size used for the simulation of the step response. After these preliminary steps one obtains an approximate transfer function $H(s)$ which may be sampled in order to apply the Loewner framework as described in Section 2.

The aforementioned procedure to determine a linear approximation of the transfer function is based on simulating the step response, i. e., using the Heaviside step function $\Theta(t)$ as input. However, we prefer to consider multiples of the Heaviside function as in [15]. For this purpose, we split the input $u(t) = a\Theta(t)$ by putting the constant factor a into the b -vector, leading to a system equivalent to (24), but replacing b by $\tilde{b} = ab$ and $u(t)$ by $\tilde{u}(t) = \Theta(t)$. Consequently, we may consider the step response without being restricted to an input magnitude of $1 W$.

We determine the step response using an input step of $10^5 W$ which is within the range of heat source magnitudes considered in [15]. The corresponding output step response is given in Figure 2. Applying the procedure outlined above, we obtain an approximation of the transfer function based on this step response. The Bode plot of this transfer function is depicted in Figure 3.

Based on the linear approximation of the transfer function, we apply the Loewner framework to obtain a low-dimensional realization which interpolates the transfer function. For this, we only need to choose interpolation points but no tangential interpolation directions, since we only have one input and one output (SISO case). As we would like to approximate the transfer function over a wide range of frequencies, logarithmically equidistant sets of interpolation points are chosen. Moreover, in order to be able to check the interpolation

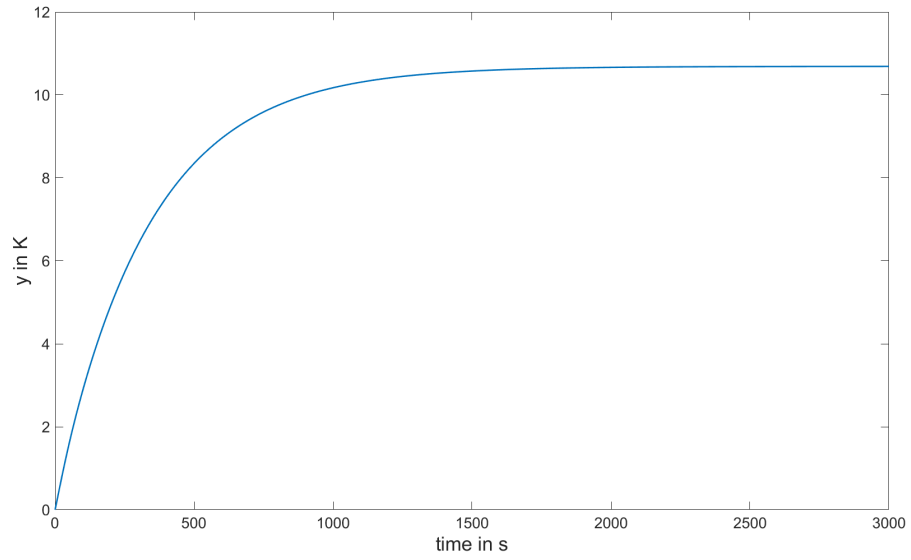


Figure 2: Step response of the full model ($u = 10^5 W$)

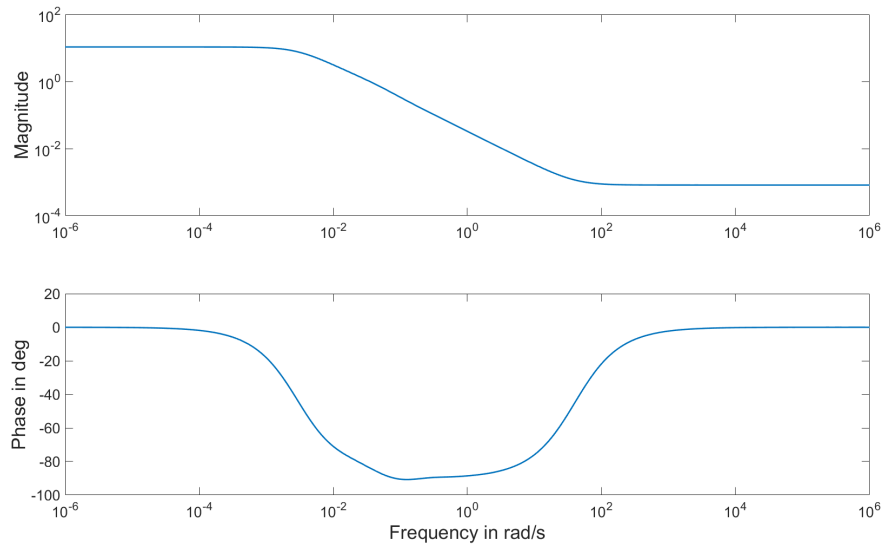


Figure 3: Bode plot of approximate transfer function

easily by means of the Bode plot, purely imaginary numbers are chosen for the interpolation points. Furthermore, the complex-conjugate interpolation points are added automatically and a coordinate transformation is performed (in function `loewner_mod`) to obtain a real-valued realization. Numerous constellations of interpolation point sets have been tested in an automatic fashion to get a better insight into proper selections of interpolation point ranges.

Comprehensive tests have shown that the range of interpolation points should not be chosen wider than 20 orders of magnitude. Ranges that are too wide lead numerically to a violation of the rank conditions which are necessary for the Loewner approach to be applicable, cf. (21). In accordance with this observation, the following rule of thumb may be formulated: The smaller the range of interpolation points, the higher the admissible number of interpolation points.

In addition to this, it should be noted that the number of interpolation points is proportional to the dimension of the Loewner realization, at least in the regular case (cf. Section 6.1). Therefore, we are mainly interested in interpolation point sets containing only a small number of points. In order to obtain real-valued realization matrices, four is the minimal number of interpolation points needed. Several constellations have been tested. The smallest step response error (measured in the maximum norm) is provided by the interpolation point set

$$S = \{-10^{3.6}i; -10^{1.8}i; -10^{-1.8}i; -10^{-3.6}i; 10^{-3.6}i; 10^{-1.8}i; 10^{1.8}i; 10^{3.6}i\}. \quad (26)$$

The dimension of the Loewner realization is half the number of interpolation points which leads to a state space dimension of four in this case. The comparison of the step response of the reduced system to that of the original system is presented in Figure 4. It should be noted that the step height for the original model is again $u = 10^5 W$, whereas the step height for the reduced model is $u = 1$, cf. discussion above about multiples of the Heaviside function. The excellent agreement of the step responses is obvious. We emphasize that a nonlinear system of dimension 15 has been reduced to a linear system of dimension four.

It is noteworthy that the reduced system only provides a good approximation for the I/O behavior of the full system, whereas the internal state variables of the original model are not captured in the reduced order model. However, in many applications, approximating the I/O behavior is sufficient, e. g., in control applications.

For the case of the interpolation point set (26), the Loewner realization is strangeness-free as well as completely controllable and observable. Thus, the index reduction and regularization procedure is not necessary in this case. In contrast, some interpolation point sets lead to realizations with strangeness-index greater than or equal to one. One example set is given by the set

$$\Sigma = \{-10^7i; -10^{-7}i; 10^{-7}i; 10^7i\}.$$

Without the index reduction procedure, the numerical integration of the resulting Loewner realization by means of the `Matlab` solver `ode15s` fails due to the higher index. However, the regularization procedure transforms the reduced system to an equivalent strangeness-free system and the numerical integration succeeds. This example emphasizes the need of a regularization procedure.

On top of potential higher-index, often unstable realizations are obtained, which are not avoided by the regularization procedure presented in this work. These systems lead to trouble when simulating the step response due to the unstable behavior. This directly leads to the topic of stability-preserving model reduction. This is not within the scope of this report but for completeness we mention the passivity-preserving interpolation approach in [1]. It also

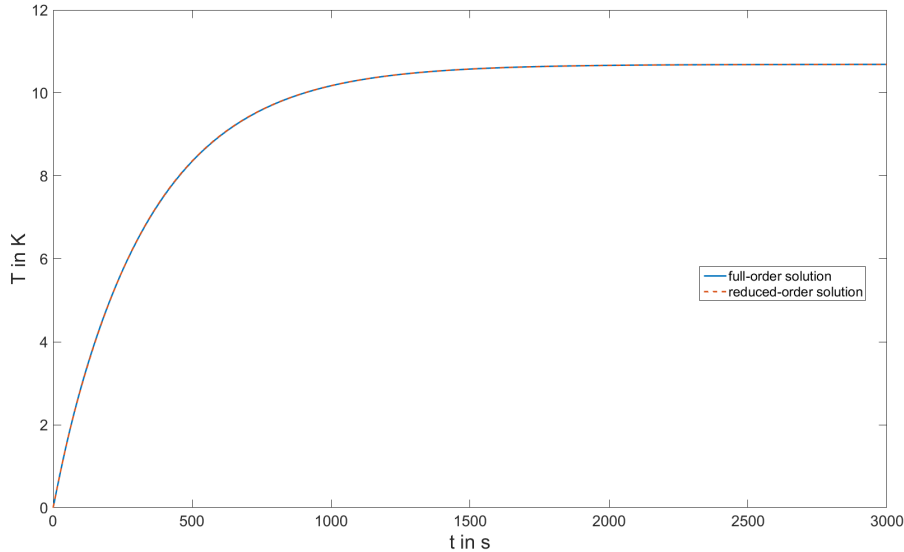


Figure 4: Comparison of step responses ($u = 10^5 W$)

preserves stability and is based on choosing the spectral zeros of the original transfer function as interpolation points. The spectral zeros are defined as the solutions of the equation

$$H(s) + H(-s) = 0.$$

As a final remark of this section, it should be emphasized that the determined transfer function and the resulting reduced order model are only valid for inputs being close to the test input $u = 10^5 W$. When considering much bigger or smaller input steps, the difference between the step responses of the reduced and the full system are significantly larger. The reason for this is the nonlinearity of the original model, which can be approximated by a linear model only locally. To illustrate this discrepancy, Figure 5 shows the comparison of the step responses for an input step of $u = 10^6 W$ where the reduced model is the same as in Figure 4 (based on step response with $u = 10^5 W$). The qualitative behavior is indeed well approximated by the reduced model but the quantitative agreement is bad when the height of the input step is much larger ($u = 10^6 W$) than that used for determining the reduced model ($u = 10^5 W$). In order to approximate the original system for a wide range of inputs, several linear surrogate models are needed or an approach different from the basic Loewner framework has to be applied.

8 Conclusion

In order to make the realization obtained from the Loewner framework suitable for simulation and control applications, we have presented a regularization procedure resulting in a strangeness-free as well as completely controllable and observable system. This procedure has been implemented in **Matlab** and is illustrated by means of a nonlinear heat transfer problem. The numerical results reveal that applying the pure Loewner realization may lead

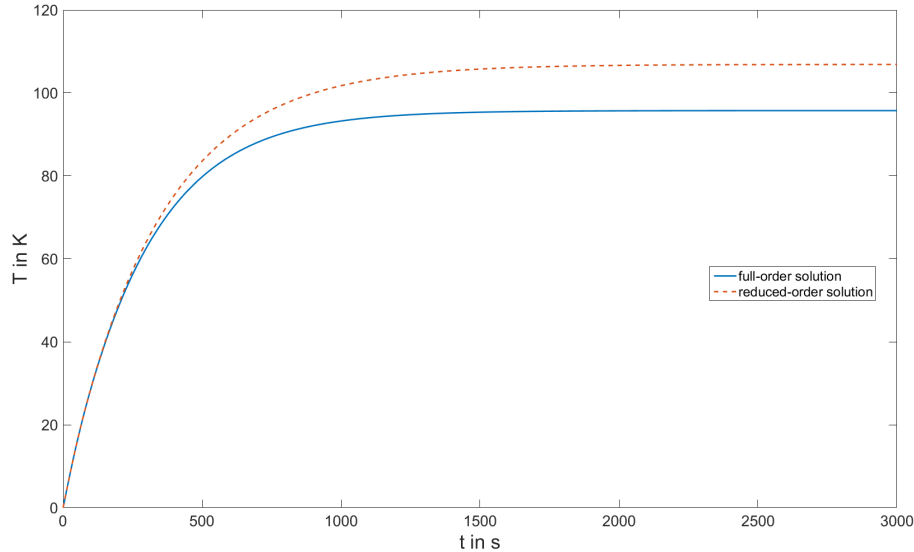


Figure 5: Comparison of step responses ($u = 10^6 W$)

to higher-index or not completely controllable or observable systems. When using the regularization procedure presented in this work, the realization is transformed to an equivalent system being strangeness-free and completely controllable and observable. These properties are important when performing simulations or when applying control methods based on the Loewner realization.

Acknowledgements. The authors gratefully acknowledge the support by the Deutsche Forschungsgemeinschaft (DFG) as part of the collaborative research center SFB 1029 *Substantial efficiency increase in gas turbines through direct use of coupled unsteady combustion and flow dynamics*, project A02 *Development of a reduced order model of pulsed detonation combustor*.

References

- [1] A. C. Antoulas. A new result on passivity preserving model reduction. *Systems & Control Letters*, 54:361–374, 2005.
- [2] P. Benner, V. Mehrmann, and D. C. Sorensen. *Dimension Reduction of Large-Scale Systems*. Springer Berlin/Heidelberg, Germany, 2005.
- [3] T. Berger. *On differential-algebraic control systems*. PhD thesis, Technische Universität Ilmenau, Germany, 2013.
- [4] J. Borggaard, E. Cliff, and S. Gugercin. Model reduction for indoor-air behavior in control design for energy-efficient buildings. In *Proceedings of the American Control Conference (ACC)*, pages 2283–2288, 2012.

- [5] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols. Feedback design for regularizing descriptor systems. *Linear Algebra and its Applications*, 299:119–151, 1999.
- [6] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols. Regularization of descriptor systems by output feedback. *IEEE Transactions on Automatic Control*, 39:1742–1748, 1994.
- [7] S. L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM Journal on Mathematical Analysis*, 18:1101–1115, 1987.
- [8] S. L. Campbell, P. Kunkel, and V. Mehrmann. Regularization of linear and nonlinear descriptor systems. In L. T. Biegler, S. L. Campbell, and V. Mehrmann, editors, *Control and Optimization with Differential-Algebraic Constraints*, pages 17–34. SIAM, Philadelphia, USA, 2012.
- [9] L. Dai. *Singular Control Systems*. Springer Berlin, Germany, 1989.
- [10] F. R. Gantmacher. *The Theory of Matrices*, volume II. Chelsea Publishing Company, New York, USA, 1959.
- [11] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, USA, fourth edition, 2013.
- [12] J. Kautsky, N. K. Nichols, and E. K. W. Chu. Robust pole assignment in singular control systems. *Linear Algebra and its Applications*, 121:9–37, 1989.
- [13] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations — Analysis and Numerical Solution*. EMS Publishing House, Zürich, Switzerland, 2006.
- [14] P. Kunkel, V. Mehrmann, and W. Rath. Analysis and numerical solution of control problems in descriptor form. *Mathematics of Control, Signals, and Systems*, 14:29–61, 2001.
- [15] J. Lienemann, A. Yousefi, and J. G. Korvink. Nonlinear heat transfer modeling. In P. Benner, V. Mehrmann, and D. C. Sorensen, editors, *Dimension Reduction of Large-Scale Systems*, pages 327–331. Springer Berlin Heidelberg, Germany, 2005.
- [16] A. J. Mayo and A. C. Antoulas. A framework for the solution of the generalized realization problem. *Linear Algebra and its Applications*, 425:634–662, 2007.
- [17] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck. *Discrete-Time Signal Processing*. Prentice Hall, Upper Saddle River, USA, second edition, 1999.